Results from a Web Impact Factor crawler[1]

Mike Thelwall
School of Computing and Information Technology
University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1SB, UK.

Email: cm1993@wlv.ac.uk

Abstract

Web Impact Factors, the proposed web equivalent of Impact Factors for journals, can be calculated by using search engines. It has been found that the results are problematic because of the variable coverage of search engines as well as their ability to give significantly different results over short periods of time. The fundamental problem is that although some search engines provide a functionality that is capable of being used for Impact calculations, this is not their primary task and therefore they do not give guarantees as to performance in this respect. In this paper, a bespoke web crawler designed specifically for the calculation of reliable WIFs is presented. This crawler was used to calculate WIFs for a number of UK universities, and the results of these calculations are discussed. The principal findings were that with certain restrictions, WIFs can be calculated reliably, but do not correlate with accepted research rankings due to the variety of material hosted on university servers. Changes to the calculations to improve the fit of the results to research rankings are proposed, but there are still inherent problems undermining the reliability of the calculation. These problems still apply if the WIF scores are taken on their own as indicator of the general impact of any area of the Internet, but with care would not apply to online journals.

INTRODUCTION

Background

Web Impact Factors (WIF) are web versions of the Impact Factors (IF) published by the Institute of Scientific Information for scientific journals. Tentative WIFs have been calculated previously with mixed results, but with some indications that they could provide some measure of the value of information in academic sites. They can be calculated, in principle, for any area of the Internet: any agreed collection of sites, or selections of pages inside sites. An area could be as large as all web pages on all sites in the national domain of a country, or it could be just a specific collection of pages inside a site, such as a directory of papers published by academics in a department. For a WIF calculation, the relevant factors are the number of pages in an area and the number of pages in another area, or collection of areas, that link to pages inside the chosen area. The WIF is then the number of pages linking to a site or area of the Internet, divided by the number of pages in that site or area [1]. A high value is presumed to indicate a site with a greater impact because there are relatively many pages linking to it.

[1] *Journal of Documentation* (2001) 57(2), 177-191

A number of studies of WIFs have been made using the advanced functionality of a search engine to compile the statistics necessary for the calculation [1,2,3]. From these it has been ascertained that the figures returned by commercial search engines are unreliable and can produce misleading results. It is known that search engine coverage of the web is incomplete and uneven [4,5] which is an additional problem, as is the fact that search engines keep secret their crawling algorithms, leaving academics to guess or deduce the rules for coverage. In order to produce reliable and well understood data, it is necessary to avoid the use of commercial search engines and to use one that can have full control exercised over it. In this paper a web crawler designed specifically for WIFs will be examined, and the results of its calculations for a set of areas on the Internet analysed.

This paper seeks to address the following questions concerning WIFs.

- Can WIFs be calculated reliably, i.e. can areas of the Internet be comprehensively crawled to enable repeatable accurate calculations?
- Can the results be reliable in the sense of being free from significant unavoidable arbitrariness?
- Does the WIF measure research impact for universities, and, if not, what does it measure?

DOCUMENTS ON THE WORLD WIDE WEB

It is important to discuss the exact nature of the World Wide Web (WWW). One definition, given in the online PC Webopaedia is:

> "*A system of Internet servers that support specially formatted documents. The documents are formatted in a language called HTML (HyperText Markup Language) that supports links to other documents, as well as graphics, audio, and video files. This means you can jump from one document to another simply by clicking on hot spots. Not all Internet servers are part of the World Wide Web*." [6]

This indicates that the Web consists solely of HTML documents. But its creator, Tim Berners-Lee conceived it differently, as *"a seamless world in which ALL information, from any source, can be accessed in a consistent and simple way*" [7]. HTML was the 'glue' that held it together, but the idea was that *all* documents could form part of the web, for example by being linked to by a HTML document. In fact, the official body of the Web, the World Wide Web Consortium, promotes a Frequently Asked Questions list which gives a vague definition of the Web, *"In practice, the web is a vast collection of interconnected documents, spanning the world*", [8]. The term World Wide Web, then, can be used either to describe either all documents accessible via a web browser, or just those written in the official language, HTML.

A web browser is a piece of software that is designed to access HTML and other documents over the web. The documents are actually transported using the Internet Protocol over the Internet, usually using the official port number of the Web, which is 80, and the hypertext transfer protocol (HTTP). An alternative definition of the web is therefore all documents obtainable over the Internet using HTTP on any port. Although the most technical of the definitions, it probably matches the impression that most users have.

From the point of view of WIFs it is important to decide which documents should be counted in WIF calculations. If an attempt is to be made to apply the calculation to unregulated parts of the web then this definition should be wide enough to encompass all common means for storing web-accessible information.

Although HTML is almost the standard format for web-based academic information, there is also another common format, Portable Document Format. This

is a file format for saving documents that can be used by magazines and Journals, and is subsequently the 'native' format for many articles. As a result many journal articles are available only as PDF documents [9,10]. Although web browsers cannot read these, there is a free reader program downloadable over the web to enable clicking on a HTML web page link to a PDF document to cause the document to be automatically loaded and displayed. PDF is not the only other format used to save web-accessible academic information and so it seems sensible to use an inclusive working definition of 'web' for WIF calculations. The chosen definition for a web-accessible document is therefore any document available over the Internet using the HyperText Transfer Protocol.

HTML AND DOCUMENT INDEXING

In order to be able to calculate WIFs, it is necessary to both count the number of web accessible documents in a domain and to identify all the links from these documents. Unfortunately both of these tasks are problematic. Web accessible documents written in HTML, web pages for short, will be discussed first. Counting the number of links from a basic web page is not a problem because there are a set of rules by which they must be declared. Counting the number of pages is, however, an issue because one page in the browser may be composed of several different HTML files. This is possible because of the frames feature, which allows one HTML document to call others to fill in rectangular areas of the browser screen. As an example an HTML file, main.htm, could instruct the browser to split the screen into two halves and to call two other files, left.htm and right.htm, to fill these areas. The user would therefore see one page although three files were used to create it. If the number of html files were counted for the WIF calculation, this would result in a score of 3, but if the number of browser screens were counted then this would score 1. Moreover, in a page designed around frames, when a link is clicked only one frame normally changes, resulting in a page which is partially the same as the previous one, and so counting screens is also misleading. The common use for frames is to have a fixed navigation bar at the top or left-hand side of the screen, whilst allowing the right hand side of the screen to change. The situation is further complicated by the fact that the individual frames of a page may or may not have been designed to be viewable on their own as a separate web page. It was therefore decided to use the simplest calculation method for WIFs and to count each file rather than each screen.

Although HTML is essentially a simple document description language, web pages are allowed to have embedded programming languages and applications as well as links to programs that automatically generate pages. These are discussed below and are normally ignored automated surveys of publicly indexable documents.

Scripting Languages

Scripting languages such as JavaScript are used in a number of web pages, possibly up to a quarter [11], because HTML is not able to interact with the browser user in any way other than processing link fetching requests. JavaScript is a programming language that can be embedded in a web page and can be activated in response to a number of user actions. One common use for JavaScript is to respond to the mouse passing over a button by changing the appearance of the button. It is, however, also capable of managing the selecting of links and the loading of new documents. It is therefore possible that a link in one page to another may not appear in the HTML, but would be activated only as a result of executing some embedded JavaScript. It was decided not to count the links in a web page that are created by an embedded scripting language, but are not in the HTML because of the practical difficulties in compiling,

executing and interpreting the results of these programs. It is believed that in nearly all cases links in JavaScript are also in the HTML, because older browsers do not support JavaScript, and therefore that this restriction is, in practice, not likely to significantly affect WIFs.

Applications

Web pages can contain information that causes applications to be executed and, usually, displayed in the browser. A simple example of this is a command that would load a video file over the internet and then cause it to be played in a rectangular area on the screen. Applications of this kind are sometimes written in the programming language Java or with Shockwave. Both of these can interact with the user to cause new screens of information to be displayed. For example a number of web pages have Java navigation buttons that act as clickable links. It was decided not to count multiple pages contained in one application or to attempt to identify application - generated links, again because of the complexity of the task. It is believed that the link count will not be greatly altered because many users disable Java in their browser for security reasons, and many others do not have Shockwave, or other applications capable of generating links, installed on their computer so that web page authors often duplicate any application generated links and information in the HTML.

Automatically created pages

Web page links can be created which send data to a program on the web server, which then selects or creates a web page in response to the actual data sent. These pages were not indexed because of being predominantly computer generated, and also because of the fact that indexing automatically generated pages which could contain links to other equally 'virtual' pages leaves open the possibility of creating an infinite loop when indexing the documents.

Server side image maps

Server-side image maps are a type of link where a picture is displayed for the user to click on and then the co-ordinates of the point clicked on are sent to the server, where a program processes the information and either directs the user to another page or automatically creates a new page. The server side image map can be implemented by its own HTML tag, or by an image button in a HTML form. It is impractical to process because it relies on the co-ordinates of a point on the image and when the user clicks on the picture, the co-ordinates of the point clicked on are sent to a program on the server, which deduces which page to send. For a map such as the UK academic clickable map, at http://www.scit.wlv.ac.uk/ukinfo/uk.map.html, its dimensions are 638 X 825 which gives 54,230 possible co-ordinates which could be sent back to the server for processing. Parsing links from this one page would therefore mean 54,230 page requests. Server side image maps are often used to provide a navigation bar, as is the case at Wolverhampton University, and so the necessary step of omitting them does have the potential to cause a problem.

Omitted pages

A number of sites have large numbers of pages that are intended for internal use only, but because the information is not sensitive, it is not protected but is left on the web. For example, Wolverhampton University has 9,998 automatically generated web pages that contain daily or monthly statistics on software and hardware attached to the internal network. These could be described by the oxymoron 'publicly accessible

Intranets'. It was decided not to include any form of automatically generated statistics, or online technical documentation (often for software or hardware) in the calculations, in addition to any areas clearly labelled as containing information for internal use only. The definition of publicly indexable for the purpose of the WIFs calculated here is therefore more restrictive than that used in other studies.

Some other areas were also excluded from the survey. Duplicate pages were not allowed. It is possible through server settings for pages or even whole areas of a site to have duplicate addresses. This often occurs as a result of moving to a new server domain name but continuing to support the old one. In some cases servers also return a different page to the one requested, for example if a protected page is requested. This can also give the appearance of a duplicate address for a page.

The web also contains experimental pages, for example one person has put over 60,000 small pages that appear to be randomly created from a dictionary source. Including these pages would significantly reduce the WIF result and so it was decided to ban any areas with over 1000 experimental pages of this type. For the same reason any teaching resources of the same size were banned. These issues will be discussed again later.

One major practical problem for a web crawler is the number of mistakes in web page links. Some are easy to identify, such as spelling a domain name incorrectly, when the mistake results in an unused or illegal domain name. Other mistakes which result in a legal address on the correct server can cause a problem if the server intelligently returns the correct address. This can produce a knock-on effect if the mistake is in the path of the address and the returned page uses relative addressing. During the testing phase this problem occurred many times, often resulting in runaway addressing. These errors were manually identified and the key initial erroneous pages added to the banned list.

THE WIF WEB CRAWLER

A web crawler program was written which possessed some of the functionality of a search engine but which produced a database containing the information required for WIF calculations. This information consisted simply of the address of each page downloaded and the addresses of all links on the page. A separate program was constructed to process the databases produced for each site or area downloaded and to calculate WIFs from them.

The method used to crawl the site was to start at the home page, and then to crawl all pages on the site linked to by the home page, and continuing to follow up all on site links in this way. This method does not guarantee to fetch all pages in an area because some may not be linked to by other pages, but does conform to the definition of *publicly indexable* [4] except it *includes* pages which are publicly visible, but requested not to appear in a search engine index and *excludes* pages on the banned list. As discussed earlier, script-generated links, application-generated links and server-side image map links were not counted, but four types of HTML links were counted

- The standard link <A HREF="page.htm">
- Client Side Image Maps <MAP> <AREA HREF="page.htm">
- Embedded Frames <FRAME SRC="page.htm">
- Automatic browser redirects such as
  <meta http-equiv=refresh content = "0;url=page.htm">

When an error occurred during and attempted retrieval of a page, this could be due to a number of factors: a temporary communications or software error; an error in the

address from the source page resulting in a non-existent (but valid) domain name; or a relatively long-term error such as a server being down. Any web page producing an error during its download was returned to later for a second attempt, unless the problem had been definitely identified as a non-existent domain name. If this attempt failed, it was recorded as incomplete and manually checked later. These pages were almost always found to be on servers that were down.

One problem that was difficult to deal with was checking whether a page downloaded was in fact the same as another page, but with a different address. This can happen in two ways: by automatic server redirection and by the existence of two copies of the document in different places. It is common for servers to allow a document to be referenced by more than one name. For example the main page of a directory can often be referenced by its name or the name of the directory without a file name. Entire directories can also have more than one valid name due to server redirection. In order to know whether a page is the same as one previously stored, it must be matched against all those previously tested. This was not possible because it required much more computing power than was available. The results will therefore include some duplication of pages. An attempt was made to manually estimate the number of pages with duplicate addresses by logging files with the same entity tag [12]. The entity tag was not used by all servers, and does not have to be unique for each file but it is rare for different files to have the same tag value. Manual checking on files with the same entity value revealed a very low rate of duplication, of under 1% of the pages crawled, indicating that the duplicate pages should not be a significant problem. The entity tag was also used to facilitate adding duplicated areas of the server to the banned list.

The sites crawled

The web crawler was pointed at the following universities to create databases for WIF calculations.

- Aston
- Birmingham
- Central England
- Coventry
- Warwick.
- Wolverhampton

These were chosen as both representative of the spectrum of UK universities and connected to the home university by an ultra high bandwidth link, an ATM connection known as MidMAN. These universities, partly as a result of this connection, might be expected to be reasonably well interconnected. The time taken to crawl each site ranged from one to five days and each site was crawled twice. Each of the sites had a standard domain name, but also ran between one and fifty-six other servers for related domain names. For example at Wolverhampton University there is a main server, on which document addresses will start with http://www.wlv.ac.uk, and a school of computing server, which is identified by http://www.scit.wlv.ac.uk. Pages were counted as belonging to each university if the domain names ended with the standard university ending, .wlv.ac.uk in the case of Wolverhampton.

The Crawling Strategy

Once the web crawler had been written and tested a banned list was compiled for each university as discussed above. It became clear during testing that it was not possible to take a snapshot of the web over a short period of time because each university had

more then one web server and at least one was often down. It would be unfair and arbitrary to survey all sites at a time when a significant proportion of one or more was not accessible. In fact some servers were not working for days at a time and some were also only up for short periods of time. The strategy adopted was to crawl a site only when all of its major servers were up and to log all pages missed on down servers. The missing pages and their links were followed at the next available opportunity. Any servers that were not found operating during the whole of the period of the survey were excluded. This whole process was repeated twice for each site and the largest database in each case was chosen for the WIF calculations.

RESULTS OF THE WIF CALCULATIONS

Table 1 shows the results of the two searches made by the web crawler. The close match between the figures supports the claim that the crawling mechanism is covering the specified proportions if the sites comprehensively. The larger differences in the case of Warwick and Wolverhampton were both found to be due to collections of pages for teaching purposes being added.

Table 1: *Page counts for two complete crawls*

Domain	Pages	Date	Pages	Date	Difference	Omitted Pages (Est.)
aston.ac.uk	51376	99-11-23	51381	99-11-29	0.0%	4,000
bham.ac.uk	84036	99-12-01	84235	99-12-03	0.2%	100,000+
coventry.ac.uk	6250	99-11-22	6253	99-11-24	0.0%	1,000
uce.ac.uk	3343	99-12-03	3349	99-12-03	0.2%	0
warwick.ac.uk	64693	99-12-08	65373	99-12-09	1.0%	40,000
wlv.ac.uk	37815	99-12-06	38163	99-12-09	0.9%	20,000

Table 2 shows a summary of the findings for the six universities covered, together with their most recent research ranking [13]. The table is in order of external link WIF.

Table 2: *Results of the WIF calculations*

Domain	**Pages**	**Self-links**	**Self-Link WIF**	**External Links**	**External Link WIF**	**Research Rank**
uce.ac.uk	3349	17821	5.32	64	0.0191	5
coventry.ac.uk	6253	22896	3.66	44	0.0070	4
warwick.ac.uk	65373	376946	5.77	405	0.0062	1
bham.ac.uk	84235	284849	3.38	485	0.0058	2
wlv.ac.uk	38163	127420	3.34	204	0.0053	6
aston.ac.uk	51381	229664	4.47	143	0.0028	3

There is clearly not a direct relationship between the WIF calculation and research ranking, supporting the findings of Smith from search engines [2]. In fact the university that scored significantly higher than the others hosts no unique research at all on its own site, only general descriptions of projects undertaken. It has scored well because of the relatively small number of pages hosted and its hosting of a local resource centre. If the list of banned pages had not been used then the rankings would have been equally unrelated to research.

The number of pages in a site

The information factor calculations for journals work on the basis that each article is attempting to provide researched information and is a potential target for references. The same is not true for web pages on a university web site. Middleton, McConnell, and Davidson [14] suggest that in addition to the staff and students of the institution there are eight groups of external users that the university may want to provide information for: prospective students; prospective staff; other academics; business people; alumni; news media; donors and benefactors; and legislators and others. It seems likely, then, that much information on any University web site would not have academic content. In addition to the pages that have been put on the web to support the function of the university, there are many that are there for other purposes, such as personal pages with biographical information and perhaps hobby details and pictures of family members. Some of the types of pages found in large numbers on the sites surveyed here are listed as follows.

- Teaching support pages
- General university information pages
- Information for prospective students
- Experimental pages
- Copies (‘mirrors') of pages based at other sites
- Online computer manuals
- Statistics of electronic events, such as web page accesses
- Student created pages for assignments
- Personal pages

All of these pages increase the denominator of the WIF calculation despite being not directly research related. The denominator is problematical because a large size in the above categories can arbitrarily reduce the WIF result, perhaps substantially. Once example of this is in automatically generated server statistics. A university such as Wolverhampton that has tens of thousands of these would be penalised, but would see their results improve if an administrator took the decision to hide the pages. There are huge numbers of teaching pages on each site, but universities with a policy of putting teaching material on the web would see their WIF score denominator increase as a result. If, however, a decision was made to password protect teaching resources, as many universities are moving to in 1999-2000, and Coventry had already done, then the WIF denominator would dramatically fall for reasons clearly unrelated to research.

Universities can clearly be ‘penalised’ for activities not directly related to research, but they can also be penalised for the format in which the information is given. For example, an online book in one large HTML page would add 1 to the WIF denominator but the same book split into chapters or sections could add tens or hundreds. The concept of a document in web terms is therefore problematical for impact calculations.

Domain sizes can also change dramatically over short periods of time for purely administrative reasons, such as clearing out old accounts or shutting down old servers. During the test period one Wolverhampton server removed all pages created by former employees in order to free disk space. The domain size is therefore likely to have periodic sudden changes.

Internal links

The Self-Link WIF shows the degree of interconnectedness of the web sites, but the necessary omission of server-side image maps has impacted upon Wolverhampton, which uses these for its standard linking tool bars on many pages.

External links

The links recorded in the survey were not just to research areas, but to many other areas of interest. The most popular pages in terms of being targets for links were:

- University home pages
- Departmental home pages
- Web navigation and information pages
- Online journals or conference home pages
- National or regional subject or resource centres
- Recreation and religion pages
- Research groups
- Teaching pages

Table 3 shows the pages with eight or more hits. Apart from the links irrelevant to research, there are also links to research not produced by the hosting institution. Online journals and conference home pages hosting the presented papers come into this category. A further example of this is the CTI Maths national subject centre had 39 links to its home page or other pages. These add to the Birmingham WIF numerator, although much of the functionality of a subject centre is not the production of new material but the hosting of that of others, and the dissemination of good practice, which, from the number of links to it, it is clearly doing. The centre also hosts links pages to subject resources, which is perhaps an example of a more fundamental problem. In the UK there are a number of initiatives to promote cross-institution collaboration, for example with respect to the broadband networks [15], and so, particularly for teaching resources, there is a widespread problem of ownership of shared resources on a server.

There is also a level of arbitrariness stemming from the level of organisation of areas of common interest. Where there is a well-organised national or international centre for teaching or research, interested academics can link to that page instead of compiling their own links pages. This can lead to well-respected pages having fewer links, because they have been linked to from the recognised source. In fact, compiling links to relevant sites of interest, including links to gateway sites, is becoming increasingly common as an important function of libraries, see [16,17,18] for example. In relation to this issue, the search engine Google uses an iterative ranking algorithm to judge the importance of documents on the web based not only on the number of pages linking to a given document, but also on the importance of the linking pages, also measured in terms of links to them [19]. This would still not differentiate between the more and less popular links if there is one recognised list of links for an academic area which attempted to be comprehensive by linking to all relevant sites.

Table 3: *Pages with eight or more external links*
A star indicates that the page no longer exists

Address	Type	Count
www.scit.wlv.ac.uk/ukinfo/uk.map.html	Navigation	70
www.bham.ac.uk/	University	34
www.uce.ac.uk/	University	27
www.bham.ac.uk/ctimath/	Centre	25
www.aston.ac.uk/	University	22
www.wlv.ac.uk/	University	21
www.coventry.ac.uk/	University	19
www.csv.warwick.ac.uk/alt-E/	Journal	15
elj.warwick.ac.uk/jilt/	Journal	14
sun1.bham.ac.uk/minnerhg/surf.htm	Journal	13
www.warwick.ac.uk/	University	12
www.cs.bham.ac.uk/	Department	11
www.cs.bham.ac.uk/system/tourist_guide/balti.html	Leisure *	10
www.cs.bham.ac.uk/~myw/fishel/	Religious *	10
clg1.bham.ac.uk/	Research	9
sun1.bham.ac.uk/m.y.zamri/msm/msmwm.htm	Religious	9
www.csv.warwick.ac.uk/	Department	9
web.bham.ac.uk/littlepa	Leisure & Research	8
www.cs.bham.ac.uk/~npa/clublist.htm	Leisure *	8
www.cs.bham.ac.uk/~npa/bars.htm	Leisure *	8
www.uce.ac.uk/tapin/tapin.htm	Centre	8

CONCLUSION

The web information factor calculation does not produce results that are closely linked to accepted research ratings primarily because of the number of web pages and links on a server that are not related to research. This is a problem that could be avoided in the long term if a convention was adopted to label all research documents or sites as such, for example as part of the Dublin Core Metadata initiative [20]. The WIF calculation could then be restricted to links to and from identified research pages. The numerator and denominator of the calculation would still be problematical, however. The denominator is problematical because of the fact that papers on the web can be presented as a single web page or a set of linked pages, and therefore one web document may or may not represent one research document. This problem could again be avoided by the use of metadata in pages to describe whether they are part of a larger document or an entire document on their own. The numerator is more of a fundamental problem. This is because in some cases a researcher wishing to link to the work of others could do this by linking to each individual document of interest, linking to a well-known search links site, or linking to the home page of a research group or hosting the pages, leading to differing link ‘accreditation’ and counts. Moreover, because most published work is still in print journals or password protected online journals, many of the research links in the study were not to individual papers, but to the home pages of researchers or research groups. For this reason also it is believed that the calculations should not just include research papers, but should also be allowed to extend to other research-related documents, even though

it is essentially a different kind of link than that when one research paper references another. It would also be necessary to exclude incoming links to online journals from the numerator of the hosting site WIF. Mirrored pages not created at the host institution would also need to be completely excluded from the calculation.

It seems therefore that there are steps that could be taken to make WIF calculations a better indicator of research potential by using metadata to clearly label web pages and by restricting the calculation to just research documents. There are two possible scopes for the calculation: either to include only online research papers or to include all research documents. The former case would be very restrictive, given that such papers are normally published elsewhere. The latter, more inclusive category, would be problematical particularly in the numerator calculation, with research areas with recognised gateway sites expected to have a smaller link count than areas where each researcher maintained their own links. In either case disciplines where computers are extensively used could be expected to have a more public web presence than others, biasing the calculation towards universities with strengths in these areas.

The widespread use of metadata to describe web pages in a way suitable for these calculations seems to be unlikely in the near future. The existing Keywords and Description metadata tags are currently used by a minority of web pages [15], with only a tiny number using the more extensive Dublin Metadata tags and so the task of promoting more complex ones seems to be a hard one.

If the proposed link of WIF calculations to research in universities is ignored then what does the calculation measure? The answer must be the general impact per page of the institution's web site over the broad range of coverage that it offers. This is still not a reliable calculation for reasons discussed before: that the number of links is not a reliable indicator of interest in a page because areas with well-organised gateway sites may well have a smaller overall number of links; and because single coherent collections of information may be on one page or several depending on the design decisions of the author.

Most of the issues discussed here would not apply to a calculation based upon online journal sites with web links to papers in other online journal sites. The only problem here would be of ensuring that single papers spanning multiple web pages were counted as one document in the calculations. A further obstacle to automatic calculations is the use of non-HTML delivery formats such as Adobe's Portable Document Format, for example for the Journal of The American Statistical Association [10], although this format does now allow the inclusion of metadata in version 4 [21]. There are also practical problems at the moment because most journals are not online, or are online but password protected. Since many journals are produced by commercial publishers, it seems unlikely that in the foreseeable future the majority of journals will be online and open access, allowing WIF calculations to be freely made.

The questions posed in the introduction to the paper have now been answered to some extent.

- WIFs can be calculated accurately, provided that a sufficiently restrictive condition is placed on the pages that are eligible to be included.
- The results are currently not free from significant arbitrariness, but with increased use of metadata the situation could be improved, but not made perfect.
- WIFs do not measure or correlate with university research profiles, they are more a measure of average interest and utility per document hosted on the site, over a wide range of types and purpose of document.

REFERENCES

1. Ingwersen, P. Web Impact Factors. *Journal of Documentation*, 54(2), 1998, 236-243.
2. Smith, A. G. A tale of two web spaces: comparing sites using Web Impact Factors, *Journal of Documentation*, 55(5), 577-592.
3. Snyder, H. and Rosenbaum, H. Can search engines be used for web-link analysis? A critical review, *Journal of Documentation*, 55(4), 1999, 375-384.
4. Lawrence, S. and Giles, C. L. Accessibility of information on the web. *Nature*, 400, 1999, 107-109.
5. Thelwall, M. Web Impact Factors and search engine coverage, *Journal of Documentation*, 56(2), 2000, 185-189.
6. PC Webopaedia, http://webopedia.internet.com/TERM/W/World_Wide_Web.html, Accessed 22 Oct 1999.
7. Berners-Lee, T. (1993). WorldWide Web Seminar. http://www.w3.org/Talks/General/Concepts.html
8. Boutell, T. (1996). What are WWW, hypertext and hypermedia? http://www.boutell.com/faq/oldfaq/htext.htm, Accessed 22 October 1999.
9. CTI-Maths. Reviews of Mathematical Software. http://www.bham.ac.uk/ctimath/reviews/, Accessed 22 Oct 1999.
10. Journal of the American Statistical Association. http://www.amstat.org/publications/jasa/index.html, Accessed 12 December 1999.
11. Thelwall, M. Commercial Web Sites: Lost in Cyberspace? University of Wolverhampton, 1999.
12. Fielding, R., Irvine, U. C., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T., Hypertext Transfer Protocol -- HTTP/1.1. ftp://ftp.isi.edu/in-notes/rfc2616.txt, June 1999, Accessed 12 December 1999.
13. Higher Education Research Rankings, http://www.thesis.co.uk/tp/999/PRN/OPEN/STATISTICS/statistics.html, Accessed 8 December 1999.
14. Middleton, I., McConnell, M. and Davidson, G. Presenting a model for the structure and content of a university World Wide Web site, *Journal of Information Science*, 25(3), 1999, 219-227.
15. Thelwall, M. Will MANs and SuperJANET dominate educational technology in the UK?, *International Journal of Educational Technology*, 1(1) 1999, http://www.amstat.org/publications/jse/.
16. Beall, J. Cataloging World Wide Web sites consisting mainly of links, *Journal of Internet Cataloging*, 1(1), 1997, 83-92.
17. Clark, J. Identifying useful Websites, *Behavioural & Social Sciences Librarian*, 17(1), 1998, 91-93.
18. Delaney, E. L. Maximising reference services in a pharmaceutical R&D library, *Electronic Library*, 17(3), 1999, 167-170.
19. Brin, S. and Page, L. The Anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 1998, 107-117.
20. The Dublin Core Metadata Initiative. http://purl.org/dc/, Accessed 12 December 1999.
21. Walter, M. Acrobat 4: Adobe's bid to make it more than a viewer, *Seybold Report on Internet Publishing*, 3(7), 1999, 3-11